Methodology.

Download the proteome sequences from the [https://wormbase.org database,](https://wormbase.org) from the base of the https://downloads.wormbase.org/species/ species found in spacie\_name/sequence/protein, downloading the most recent .protein.fa file.gz at the time of consultation. The species are as follows, along with relevant information about their sequence.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Indicator | Name | Sequence Version | Sequence date | Size |
| PRJNA231479 | *Ancylostoma ceylanicum* | WS248 | 3/25/2015 | 7.4 M |
| PRJNA62057 | *Ascaris suum* | WS248 | 3/25/2015 | 3.6 M |
| PRJNA10729 | *Malayi witch* | WS286 | 08/09/2022 | 3.5 M |
| PRJEA64437 | *Bursaphelenchus xylophilu* | WS248 | 3/25/2015 | 3.8M |
| PRJNA51225 | *Caenorhabditis angaria* | WS286 | 08/09/2022 | 4.9 M |
| PRJEB28243 | *Caenorhabditis becei* | WS286 | 08/09/2022 | 5.8M |
| PRJEB34497 | *Caenorhabditis bovis* | WS286 | 08/09/2022 | 3.8M |
| PRJNA20035 | *Caenorhabditis brenneri* | WS286 | 08/09/2022 | 7.6M |
| PRJNA10731 | *Caenorhabditis briggsae* | WS286 | 08/09/2022 | 5.5M |
| PRJEB28388 | *Caenorhabditis elegans* | WS286 | 08/09/2022 | 6.1M |
| PRJDB5687 | *Caenorhabditis inopinata* | WS286 | 08/09/2022 | 5M |
| PRJNA12591 | *Caenorhabditis japonica* | WS286 | 08/09/2022 | 5.6M |
| PRJNA248912 | *Caenorhabditis latens* | WS286 | 08/09/2022 | 5.7M |
| PRJNA384657 | *Caenorhabditis nigoni* | WS286 | 08/09/2022 | 7.2M |
| PRJEB28259 | *Caenorhabditis panamenesis* | WS286 | 08/09/2022 | 5.6M |
| PRJEB12595 | *Caenorhabditis parvicauda* | WS286 | 08/09/2022 | 3.9M |
| PRJEB11354 | *Caenorhabditis quiockensis* | WS286 | 08/09/2022 | 5.1M |
| PRJNA53967 | *Caenorhabditis remanei* | WS286 | 08/09/2022 | 7.5M |
| PRJNA194557 | *Caenorhabditis sinica* | WS286 | 08/09/2022 | 8.5M |
| PRJEB12601 | *Caenorhabditis sulstoni* | WS286 | 08/09/2022 | 4.7 M |
| PRJEB12608 | *Caenorhabditis tribulationis* | WS286 | 08/09/2022 | 5.9M |
| PRJNA53597 | *Caenorhabditis tropicalis* | WS286 | 08/09/2022 | 5.3M |
| PRJEB12600 | *Caenorhabditis uteleia* | WS286 | 08/09/2022 | 6.5M |
| PRJEB12602 | *Caenorhabditis waitukubuli* | WS286 | 08/09/2022 | 6.2M |
| PRJEB12596 | *Caenorhabditis zanzibari* | WS286 | 08/09/2022 | 5.4M |
| PRJEB1797 | *Dirofilaria immitis* | WS248 | 3/25/2015 | 3M |
|  | *Ditylenchus dipsaci* |  |  |  |
| PRJNA19377 | *Heterorhabditis bacteriophora* | WS248 | 3/25/2015 | 2.8M |
| PRJEB506 | *Haemonchus contortus* | WS248 | 3/25/2015 | 5.1M |
| PRJNA60051 | *Loa loa* | WS248 | 3/25/2015 | 3.1M |
| PRJNA29083 | *Meloidogyne hapla* | WS248 | 3/25/2015 | 3.0M |
| PRJNA28837 | *Meloidogyne incognita* | WS248 | 3/25/2015 | 4.1M |
| PRJNA72135 | *Necator americanus* | WS248 | 3/25/2015 | 3.2M |
| PRJEB15512 | *Oscheius tipulae* | WS286 | 08/09/2022 | 4M |
| PRJEB513 | *Onchocerca volvulus* | WS286 | 08/09/2022 | 3.2M |
| PRJEB6009 | *Pristionchus exspectatus* | WS248 | 3/25/2015 | 5.4M |
| PRJNA12644 | *Pristionchus pacificus* | WS286 | 08/09/2022 | 6.9M |
| PRJNA186477 | *Panagrellus redivivus* | WS286 | 08/09/2022 | 5.5M |
| PRJEB125 | *Strongyloides ratti* | WS286 | 08/09/2022 | 3.8M |
| PRJEB126 | *Trichuris muris* | WS286 | 08/09/2022 | 3.8M |
| PRJNA12603 | *Tichinella spiralis* | WS236 | 02/12/2013 | 3.1M |
| PRJNA208416 | *Trichuris suis* | WS248 | 3/25/2015 | 3.1M |

To obtain the proteases, the MEROPS database of page <https://www.ebi.ac.uk/merops/download_list.shtml> is used, specifically the complete sequence peptidase library was used. "4. Peptidase Full-length Sequences". It is downloaded using the command "sudo ftp https://ftp.ebi.ac.uk/pub/databases/merops/current\_release/protease.lib", using the linux subsystem (with Ubuntu) for Windows 11. Later the protease.lib library is indexed to be able to make a local blast.

For this you must first install the blast+ package using the command "sudo apt-get install ncbi-blast+"

Then to index and prepare the portease.lib library use the command "makeblastdb -in protease.lib -dbtype "prot" -out protease"

Subsequently, the local blasts are performed by

blastp -query Nematodes/species/species/proteina.fa -db database\_protease/protease -num\_threads "8" -max\_target\_seqs "2" -evalue "0.01" -outfmt "6" > blast\_.txt

The -num\_threads option "8" indicates that 8 cores are used for the operation.

The option -max\_target\_seqs "2" tells us that alone we reroten the best 2 matches

The -evalue option "0.01" tells us that we choose the e-value at 0.01

The option of outfmt "6" gives us as an option a tabular blast file with the following columns.

Of the blasts obtained short the first 3 columns with cut -f 1-3 "$file" > "${file%.txt}\_cut.txt to obtain only the columns with the indicator of the protein, indicator of the protease and percentage of intensity between the protein and the protease

For the quantification of how many proteases of each proteome

First we must create a table that relates the MER indicators ##### with the name of the protease. For this we go to the protease.lib file of the protease database and run the "create\_protease\_index.py" script

Then you have to run the script "count\_protease.py" to generate a .csv file called summary\_table which contains the amount of proteases of each proteome depending on the type of family and subfamily, considering the following table:

|  |  |
| --- | --- |
| Portease type | Family/Subfamily |
| S | Serine Pepidases |
| C | Cysteine Peptidases |
| To | Aspartic Peptidases |
| M | Metallo Peptidases |
| I | Inhibitors |
| T | Threonine Peptidases |
| G | Glutamic Peptidases |

The following considerations must be taken into account when running the script.py:

* In the def main() function:

Blast\_folder = Place the address where the blast files are stored

indicator\_file = Place the address where the indicators\_without\_source.txt file created by the create\_protease\_index.py script is stored, preferably in a directory other than blast\_folder to avoid indexing errors

blast\_output\_folder = "Blast\_complete\_indicator", this Directory is created by the script and stores the blasts but with the flag MER##### replaced by its extended correspondence of indicators\_without\_source.txt

best\_match\_output\_folder = "blast\_best\_match" In this directory that creates the script, the blast files are saved, but only with the best coindication for each sequence.

protease\_count\_output\_folder = "protease\_count" this Directory that creates the script an arhcivo .txt of each species is saved in which there are how many coincidences there are that each indicator MER#####

renamed\_output\_folder = "blast\_best\_match/count/shortened\_name", in this Directory that creates the script the protease\_count files are saved with the name shortened to only the name of the species.

summary\_output\_file = "summary\_table.csv" output name of the final file.